Unsupervised Non-topical Classification of Documents

Ron Bekkerman RONB@CS.UMASS.EDU
Department of Computer Science
University of Massachusetts, Amherst, USA

Koji Eguchi EGUCHI@NII.AC.JP
Kobe University, Kobe, Japan, and
National Institute of Informatics, Tokyo, Japan

James Allan ALLAN@CS.UMASS.EDU
Department of Computer Science
University of Massachusetts, Amherst, USA

UMass CIIR Technical Report IR-472

Abstract

We describe the problem of non-topical clustering of documents, the purpose of which is to divide a set of documents into clusters that share some aspect. We present experiments on the British National Corpus that cluster documents by genre. We show that words are superior to part of speech information for genre clustering, but that better results can be obtained by using both. We also demonstrate that the new multi-way distributional clustering approach is highly effective for this task because it requires less feature crafting than other techniques.

1. Introduction

Document clustering by topic is very well known [11] and has a lengthy history within the field of information retrieval [31]. Clustering has been used to improve the efficiency of retrieval, to provide a form of query expansion (by drawing in similar documents that might not share the query terms), and to group the the retrieved documents into sets that are on roughly the same topic. It is this last goal of clustering that interests us: the goal of helping the searcher get a sense of what the retrieved set was about without having to read as many documents. When clusters are well described, the searcher can recognize the set of topics and select the ones that are of value.

Topics, however, are not the only way in which someone might want to select groups of documents. Aspects such as genre, opinion, authorship, style, mood, and so on are interesting dimensions along which retrieval results might break. For example, someone might be interested in upbeat documents on a topic, or fictional accounts of an event. In this research, we focus on techniques appropriate for such non-topical grouping, with a particular emphasis on genre.

Two fundamental ways in which documents can be grouped by topic or by genre are supervised or unsupervised. In the former, a set of training documents is appropriately classified and a learning algorithm is used to find features that are indicative of the groups.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Unsupervised Non-topical Classification of Documents	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts Amherst,Department of Computer Science,140 Governors Drive,Amherst,MA,01003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We describe the problem of non-topical clustering of documents, the purpose of which is to divide a set of documents into clusters that share some aspect. We present experiments on the British National Corpus that cluster documents by genre. We show that words are superior to part of speech information for genre clustering, but that better results can be obtained by using both. We also demonstrate that the new multi-way distributional clustering approach is highly effective for this task because it requires less feature crafting than other techniques.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**16**	

Provided they are similar in nature to the training data, new documents can then be appropriately classified.

When the classification is needed on a corpus unlikely to have any training data, or along a dimension for which there is no existing classification, a system can fall back on unsupervised approaches, i.e., clustering. Such use of clustering will leverage human experience to select the general types of features (e.g., words, parts of speech, dates, punctuation) that are likely to distinguish between groups, but will have little or no explicit training data.

The focus of this research is (unsupervised) document clustering by genre. Although the field of non-topical (supervised) classification is well explored in the literature (a lot of work has been done on classification by genre [12, 13, 9, 17, 26], by text authorship [18, 1], by writer's gender [14], tone [30, 22] and mood [21], as well as by familiarity with the topic of the discussion [15]), we believe that this paper is the first comprehensive study of the problem of genre *clustering*.

In addition to helping a searcher identify documents of the right type, we envision other uses of this genre clustering: (a) an automatic assistant to a librarian that provides preliminary grouping of documents by genre; (b) a component of a system that provides preliminary grouping of training data for genre *classification*; (c) a routing framework for situations where an algorithm is known to work differently on different genre (e.g., summarization, segmentation), but for which there is no training data broken down by genre.

In the next section we provide a formal problem statement of topical and non-topical clustering and in Section 3 we describe work related to both genre classification and clustering. We discuss our approach to representing documents as well as our clustering and evaluation methods in Section 4. In Section 5 we describe how we used the British National Corpus for our experiments. The results are presented and discussed in Section 6. In Section 7 we conclude and outline our future work.

2. Problem statement

Throughout this paper by "clustering" we mean "hard clustering", which is a problem of data partitioning where a data point is assigned to one cluster only, while in "soft clustering" each data point is represented as a distribution over the resulting clusters. Formally, given a set X of n data points, the hard clustering problem is to create $k \ll n$ subsets $\{X_i\}_{i=1}^{k}$, such that they cover the entire set: $X = \bigcup_{i=1}^{k} X_i$, and the subsets do not intersect: $\forall i \neq j \; X_i \cap X_j = \emptyset$. The constructed subsets X_i are useful only if they contain data points that are "similar" to each other according to a certain criterion, e.g. they are close in a Euclidian space. It is also desirable that data points assigned to different clusters would be less similar to each other than the data points within one cluster.

From the probabilistic perspective, the hard clustering problem is defined as follows. A discrete random variable X is defined over the set X and the goal of a clustering method is to construct a random variable $\tilde{X}$ with values $\{\tilde{x}_i\}_{i=1}^{k}$ such that $\forall x_i \; \exists \tilde{x}_j : \; P(\tilde{x}_j|x_i) = 1$. In Machine Learning, clustering is often used for density estimation of the variable X, i.e. for generating a "summary" of the data behavior. In Information Retrieval, data clustering often plays the role of dimensionality reduction. In particular, feature clustering aims at constructing a more compact, yet meaningful representation of documents (see, e.g., [5]).

Obviously, density estimation and dimensionality reduction are in fact two aspects of the same problem of compact data representation.

There are two main competitive approaches for clustering data: a vector space approach and a Bayesian approach. Most clustering algorithms to-date can be assigned into one of these classes. In a vector space model data points are represented as vectors of features and a distance measure is defined for these vectors. Close vectors are then put into the same cluster. One of the instances of the vector space model is distributional clustering [2]. In distributional clustering, each value of the random variable X is represented as a distribution over another, correlated random variable Y. The goal is to cluster these distributions such that similar distributions are located together in one cluster. A generalization of this method is Information Bottleneck [29], in which distances between the data points are not explicitly computed but a clustering $\tilde{X}$ of size k is built that *maximizes information* about Y. This is expressed in terms of Mutual Information $I(\tilde{X};Y)$ (see Equation 2).

Bayesian clustering methods usually employ a *generative* process: a directed graphical model is proposed that prescribes rules according to which the given data has presumably been generated. The model should in some way incorporate the idea that the data came from a source where it had already been clustered and now this clustering should be revealed. Parameters of the model are then estimated on the data and the posterior distribution of the data over the clusters is calculated. This usually produces soft clustering that can be then "hardened" by considering only the most probable cluster for each data instance.

In this paper we discuss two state-of-the-art clustering methods (see Section 4.2): one of the vector space family (Multi-way Distributional Clustering [4]) and another one of the Bayesian family (Latent Dirichlet allocation [7]). We compare the performance of these two methods on the problem of document clustering by genre.

Clustering by genre can be broken into basically the same components as clustering by topic. We first choose an appropriate document representation on which we then apply our clustering algorithms. Although the algorithm application appears to be identical for both tasks, the document representation should supposedly be different.

As the task of clustering by topic is well explored, most researchers use the simplest document representation *Bag-Of-Words (BOW)*, assuming that words carry most of a document's content. This assumption is not always true, though. Other potentially useful features can be taken into account, such as punctuation, markup, authors' names etc. Some researchers study richer document representation models [4, 24], but BOW remains most popular because it usually produces surprisingly good results (for some discussion, see [3]).

However, a common intuition tells that the BOW representation is not appropriate for *non-topical* (supervised or unsupervised) classification. Probably, the reason for this belief is that BOW fits well into the topical classification task. If we are convinced that *the topic is in the words*, then we assume that the *"non-topic"* (style, genre, mood etc.) should not be in the words. Logically speaking, this might be wrong. Words may carry additional information. One of the goals of this paper is to check this assumption. For clustering by genre, we represent documents both as bags of words (i.e. bags of "contextual" features) and as bags of part-of-speech ngrams (which are believed to carry stylistic information). We compare clustering results on both representations. We also *combine* both representations, exploiting the fact that our multi-way distributional clustering model allows a straightfor-

ward combination of various document representations. We suppose that such combination would lead to cleaner clustering.

3. Related Work

Swales [28] defines a genre as "a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale of the genre" (p.58). Lee [16] clarifies this definition as well as other related terms: he describes the genre as a category assigned on the basis of external criteria such as intended audience, purpose and activity type. In this paper, we follow Lee's definition. We also assume that there is a significant correlation between the genre and a grouping of documents written in a similar style, and thus linguistic characteristics of the text play an important role in distinguishing between the genres, as suggested in [12, 13, 9, 17, 26].

While most of the related work is on *supervised* non-topical classification, we have found two manuscripts in which unsupervised methods were considered. A pioneering study was conducted by Douglas Biber in the late eighties [6]. He attempts to automatically identify *text types*, which refer to groups of documents according to their linguistic content (e.g., `informational production`, `narrative concern` etc.), irrespectively of their genre categories. Biber applies the *multi-dimensional analysis* method using patterns of manually detected linguistic features, such as tense, aspect markers and anaphora. In contrast to Biber's work, we cluster documents by genre, and automatically construct highly discriminative features using the state-of-the-art multi-way distributional clustering model.

The work of Rauber et al. [23] is most closely related to ours. They perform genre clustering of documents organized according to a certain topic, using domain independent features such as frequencies of special characters, punctuation and stopwords. They apply "self-organizing maps", a neural network learning model, for clustering the feature vectors. The goal of Rauber et al.'s work is to incorporate genres into the topic-based organization of a digital library. Genre clustering is performed only on topically coherent groups of documents. No comprehensive study of the nature of document clustering by genre is conducted. We focus exclusively on document clustering by genre, and evaluate it on a collection of both topically and stylistically heterogeneous documents, while testing a variety of clustering methods and document representations.

Lee et al. [17] perform sophisticated feature selection in the context of supervised genre classification. Their method is based on identifying the terms that occur in many documents of a certain genre while being uniformly distributed over topical classes, assuming that the genre-revealing terms should be independent of the topic. In their work, only the the Bag-Of-Words model is used. We also assume in our paper that the Bag-Of-Words model is effective for discriminating between genres, especially when used together with stylistic features such as parts-of-speech and punctuation. Rather than performing feature selection, we focus on feature construction using a multi-way clustering method.

Argamon et al. [1] study the distributions of unigrams, bigrams and trigrams of parts-of-speech, as well as pronouns and determiners, in the BNC corpus[1] and disclose significant differences between non-fiction and fiction documents and between author genders. Follow-

1. We also use the BNC corpus for evaluation of our methods, see Section 5.

ing this work, Santini [25] uses uni-/bi-/trigrams of parts-of-speech with or without punctuation for a supervised genre classification task on the BNC corpus. As we will discuss later in this paper, the part-of-speech ngram model is not the best model for distinguishing genres in the BNC corpus. We will compare the two document representations (bag-of-words and part-of-speech ngrams) for the task of unsupervised genre classification.

4. Methods

4.1 Document representation

In this work we explore two types of document representation: Bag-Of-Words (BOW) and a bag of ngrams of Part-Of-Speech tags. We make the following observation which might seem counterintuitive at the first glance: Bag-Of-Words is a perfectly appropriate document representation for the problem of genre classification / clustering, because vocabularies that are used in different genres are significantly different. For example, the word "retrieval" appears in our dataset in 22 documents, but never in fiction.

Tuples of POS tags presumably reveal stylistic characteristics of documents. It is commonly believed, e.g., that passive voice constructions are more often used in formal discussion of the scientific or juristic literature, rather than in fiction or news stories. Such phenomena are captured by sequences of POS tags. We define an ngram as a sequence of n tags extracted out of a valid sentence. The ngrams are extracted from the sentence in an incremental manner: the first ngram starts with the tag of the first word in the sentence, the second one starts with the tag of the second word etc. The last ngram ends with the tag of the last word of the sentence. For example, out of the sentence

```
<PNP>It <VBZ>'s <AT0>a <AJ0>real <NN1>holiday <PUN>.
```

we extract four trigrams:

```
PNP_VBZ_AT0, VBZ_AT0_AJ0, AT0_AJ0_NN1, AJ0_NN1_PUN.
```

All the ngrams extracted from a document are then stored in a term frequency vector in complete analogy to the Bag-Of-Words representation.

4.2 Clustering methods

In this section we discuss two powerful clustering methods that we apply to the problem of unsupervised genre classification.

4.2.1 Multi-way Distributional Clustering

Bekkerman et al. recently introduced the *multi-way distributional clustering (MDC)* [4], which is an information-theoretic clustering scheme whose origins come from the Information Bottleneck [29]. The power of the method is in *simultaneously* constructing N clusterings of N random variables defined over a given dataset, while exploiting pairwise interactions between the variables. In the text domain, such variables can be defined over documents, their words, author names, titles etc.

Let us first discuss the motivation for performing the simultaneous clustering procedure. Many (if not most) of the vector-space text classification / text clustering systems involve

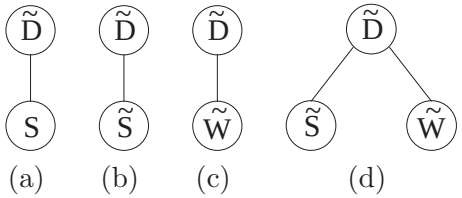

Figure 1: Pairwise interaction graphs for: (a) 1-way MDC with POS unigrams; (b) 2-way MDC with POS 2-grams, 3-grams or 4-grams; (c) 2-way MDC with BOW; (d) 3-way MDC with POS and BOW.

some form of feature selection as a preprocessing step. At this step, the practitioner decides what type of features would be most appropriate for solving the current problem, after which the documents are represented as vectors of the extracted features. Legitimateness of the features is then empirically evaluated. An obvious problem of this approach is that it is unclear whether there exists *another* set of features that would better match the desired goal. A method for constructing the features *on the spot*, while the document classification / clustering is performed, would be desirable. And since one of the most powerful feature construction techniques is feature clustering (see, e.g., [5]), the simultaneous clustering of documents and their features would be the best choice.

Simultaneous clustering methods have recently emerged in many machine learning fields including bioinformatics, machine vision and collaborative filtering as well as in text clustering (for a short survey, see [4]). Most of these methods perform two-way clustering, but MDC is capable of simultaneously clustering $N > 2$ variables, while staying within a reasonable computational complexity.

Given a particular textual dataset, let D be a random variable over its documents, W be a random variable over its words and S be a random variable over the Part-Of-Speech ngrams of its words. The goal is to construct a clustering $\tilde{D}$ of documents D, while simultaneously constructing a clustering $\tilde{W}$ of words W and/or a clustering $\tilde{S}$ of POS ngrams S, by maximizing the pairwise Mutual Information between interacting variables.

In MDC, relevant interactions between the variables are represented in an *pairwise interaction graph* $G = (X, E)$, where vertices $X = \{X_1, \ldots, X_m\}$ are the random variables occurring in the system, and edges E connect pairs of variables whose Mutual Information we want to maximize. Examples of pairwise interaction graphs are given in Figure 1. Our objective function is then

$$\max_{\{X_i\}} \sum_{(X_i, X_j) \in E} I(X_i; X_j), \tag{1}$$

where $I(X_i; X_j)$ is the Mutual Information

$$I(X_i; X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}. \tag{2}$$

In this paper, we consider four practical cases of MDC models:

1. **POS unigrams.** Since the number of different POS tags in any tagging system is relatively small, it makes no sense to simultaneously cluster documents and POS unigrams. Therefore, we apply a 1-way MDC model: given a pairwise interaction graph from Figure 1(a), we maximize the objective function from Equation (1), which in this simple case has the form of $I(\tilde{D}; S)$.

2. **POS ngrams, where $\mathbf{n > 1}$.** The number of unique POS ngrams of order higher than 1 is exponential in the number of POS tags, so their clustering is desired. We perform a 2-way MDC with the pairwise interaction graph from Figure 1(b) and the objective $I(\tilde{D}; \tilde{S})$.

3. **Bag-Of-Words.** The number of unique words in a dataset is comparable with the number of POS trigrams, so in analogy to the previous model, we perform a 2-way MDC with the pairwise interaction graph of Figure 1(c) and the objective $I(\tilde{D}; \tilde{W})$.

4. **BOW+POS hybrid.** We combine contextual information of BOW and stylistic information of POS ngrams into a 3-way MDC model, where we simultaneously cluster documents, words and bigrams of POS tags. Given the pairwise interaction graph of Figure 1(d), we maximize the sum $I(\tilde{D}; \tilde{S}) + I(\tilde{D}; \tilde{W})$. Note that we do not consider the interaction between $\tilde{S}$ and $\tilde{W}$ because these variables are almost coupled. Words can be clustered according to their POS tags totally *regardless* of their distribution over documents, while we are interested in word clusters that would shed some light on the implicit structure of the particular document collection.

Note that in all the cases listed above we have to fix the number of clusters $|\tilde{D}| = k_d$, $|\tilde{S}| = k_s$ and $|\tilde{W}| = k_w$ while maximizing the objective function, in order to avoid a degenerative partitioning. This does not imply, though, that we are restricted to using only flat, k-means-like clustering algorithms. On the contrary, we can fully exploit the hierarchical nature of clusters, in the following manner: starting with any initial configuration, we can split or merge clusters randomly, or according to a certain criterion, and then we fix the current number of clusters and perform a *correction procedure* in which we rearrange elements within the clusters while maximizing the objective function.

In the top-down clustering scheme, we start with the configuration where all the elements are placed in one cluster. We then split the cluster into two halves and apply the correction procedure: greedily, we remove each element from its cluster and try to put it into any other cluster while recalculating the objective; we then leave it in the cluster where the objective achieves its local maximum. This setup guarantees the convergence of our optimization method. When no element can be further relocated, we stop the correction procedure, and are now allowed to split the clusters again. Analogously, in the bottom-up scheme we start with singleton clusters and merge each cluster with its closest peer, after which we perform the same correction procedure as for the top-down scheme, and so on. We apply these schema iteratively: at each iteration we select a node of the pairwise interaction graph and optimize it, then select another node and optimize it, in a round-robin fashion.

Bekkerman et al. [4] show that in order to achieve the optimal computational complexity of the MDC algorithm, the bottom-up scheme should be applied to one of the variables in the system, and the top-down scheme to all the other variables. Our bottom-up variable is the document clustering $\tilde{D}$. For other details on the MDC clustering algorithm, see [4].

4.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [7] is a Bayesian generative model that has recently received much attention in the machine learning community (see, e.g. [24, 19]). Blei et al. [7] define *topics* of documents as multinomial distributions over words in the dataset. Since a document can potentially discuss a number of topics, for each document a distribution over the topics can be defined. A Dirichlet random variable is then defined over these distributions. Blei et al. use the variational approximation technique to compute a posterior Dirichlet distribution of each document over the topics. In our paper we estimate this posterior using the Gibbs sampling inference technique (as described in [24]): for each document we first sample a distribution over the topics from the Dirichlet distribution. Then for each position of a word in the document we sample a particular topic from the chosen distribution after which we sample a specific word from the chosen topic.

A formal description of the LDA method introduces a lot of new notation and is beyond the scope of this paper.

4.3 Evaluation methodology

Numerous methods for clustering evaluation have been proposed during the last century (for a survey, see [11, 20]), most of which fall into one of the two groups: some of them measure the homogeneity of clusters based on intrinsic criteria, while the other compare the clusters with some type of ground truth. We believe that methods of the second group are more appropriate for our current work. Despite the large variety of such methods, we are not familiar with any method that would have no drawbacks. Therefore, we choose two evaluation criteria that are not perfect but in our opinion are most straightforward and intuitive: these are *micro-averaged accuracy* and *Jaccard index*:

- **Micro-averaged accuracy.** Let X be the target variable and $\tilde{X}$ its clustering. Let C be the set of ground truth categories. For each cluster $\tilde{x}$, let $\gamma_C(\tilde{x})$ be the maximal number of $\tilde{x}$'s elements that belong to one category. Then, the precision $Prec(\tilde{x}, C)$ of $\tilde{x}$ with respect to C, is defined as $Prec(\tilde{x}, C) = \gamma_C(\tilde{x})/|\tilde{x}|$. The micro-averaged precision of the entire clustering $\tilde{X}$ is:

$$Prec(\tilde{X}, C) = \frac{\sum_{\tilde{x}} \gamma_C(\tilde{x})}{\sum_{\tilde{x}} |\tilde{x}|}, \tag{3}$$

  that is, the portion of documents appearing in the dominant categories. Note that this measure is meaningless when the number of clusters $|\tilde{X}|$ is large. In particular, if $|\tilde{X}|$ equals the number of data points, the micro-averaged precision is 1. Actually, this measure makes sense only when $|\tilde{X}|$ equals the number of categories. Then $Prec(\tilde{X}, C)$ equals the standard *micro-averaged accuracy*. Note that the micro-averaged accuracy is in fact a compliment to 1 of another popular evaluation measure, called *classification error*.

- **Jaccard index** compares sets of clusters and ground truth categories at the level of document *pairs*. Let a be the number of document pairs that belong to the same category and are placed into the same cluster. Let b be the number of document pairs that belong to the same category but are found in different clusters. Let c be the

number of document pairs that happen to fall into the same cluster but belong to different categories. The Jaccard index [10] is then:

$$J(\tilde{X}, C) = \frac{a}{a+b+c}.$$

The index equals 1 for the perfect match of clusters and categories. Note that since the Jaccard index in defined on *pairs* of documents, its dimensionality is actually quadratic in the number of documents. It would be more appealing then to take a square root of $J(\tilde{X}, C)$, but we will stick to its standard form as presented above.

The major advantage of the accuracy measure is that it takes into account only the dominant category of each cluster which naturally corresponds to the human evaluation: when looking at a cluster, a person would determine the cluster's topic according to the topic of the *majority* of the cluster's members, while considering the rest of the cluster as noise with respect to the dominant topic. The Jaccard index, however, treats all the categories inside each cluster equally, which is an obvious disadvantage.

The main drawback of the micro-averaged accuracy is that it does not penalize a split of a category over a number of clusters, as long as the category remains dominant in the clusters. For example, given three categories $\{c_1, c_2, c_3\}$ and three clusters, if category c_1 dominates in clusters 1 and 2, while category c_2 dominates in cluster 3, the accuracy is not hurt by the fact that category c_1 is split over two clusters, while category c_3 is totally excluded from the accuracy calculation. In contrast, the Jaccard index takes c_3 into account, and penalizes the split of c_1 (fewer *pairs* of documents fall into the same cluster).

In all our experiments, we fix the number of document clusters to the actual number of categories. Since our algorithms are randomized, we report on *average* micro-averaged accuracy and *average* Jaccard index, taken over four independent runs.

5. Dataset

For evaluation of our methods, we use the British National Corpus (BNC) [8]. The corpus consists of 4054 texts of written and spoken English language. In this work we consider only the written part of the BNC (3144 documents). The original BNC is not annotated by genre. Following Santini [25] we use David Lee's ontology of BNC genres [16]. The ontology is extremely fine-grained: it consists of 46 genres that cover most aspects of the modern literature: fiction prose and poetry, non-fiction academic and non-academic texts, national and local newspaper articles, religious texts, advertisements etc. We exclude categories whose titles contain words "other" or "misc" assuming that their content is presumably too vague.

As mentioned in Section 4.3, our evaluation measure (the micro-averaged accuracy) takes into account only the largest category of each cluster. It is natural then that if a dataset has a few categories that are dominant in size over the rest of the categories, documents of the those categories would probably prevail in the clusters. Therefore, the fair evaluation using the clustering accuracy would be performed on a dataset that contains categories of similar size. However, it is not the case in BNC. Some of its categories are very large (e.g. 432 documents in `W_fict_prose`), while some of them are extremely small (only 2 documents in `W_fict_drama`). First, we exclude categories that are not composed

Document representation	k-means		LDA		MDC	
	Accuracy	Jaccard	Accuracy	Jaccard	Accuracy	Jaccard
POS bigrams	0.232	0.089	0.447 ± 0.002	0.185 ± 0.001	0.510 ± 0.002	0.224 ± 0.001
Bag-Of-Words	0.091	0.046	0.554 ± 0.001	0.259 ± 0.001	0.557 ± 0.002	0.256 ± 0.002
BOW + POS bigrams					$\mathbf{0.585 \pm 0.006}$	$\mathbf{0.274 \pm 0.005}$

Table 1: Clustering results, averaged over four independent runs. Standard error of the mean is shown after the ± sign.

of a representative number of documents (less than 32 documents). Second, for each of the remaining 21 categories, we uniformly at random choose 32 documents, so our resulting dataset consists of 672 documents. A list of the 21 genres can be found in the first column of Table 2.

The BNC texts are represented in SGML. We remove all the markup, remaining with the pure text only (located between the `<body>` and `</body>` tags). We do not stem, but do lower case of letters. The preprocessed corpus contains 169590 unique words, half of which appear only in one document. All words in the British National Corpus are semi-manually tagged according to their Part-Of-Speech (POS). The POS tagging system consists of 91 tags, 30 of which are ambiguous (such as `AJ0-VVD`, which means that it is unclear whether the word is an adjective or a past tense verb). Four of the tags (`PUL`, `PUN`, `PUQ` and `PUR`) refer to the punctuation.

In all our experiments we ignore low frequency words and POS ngrams (the ones that appear only in one or two documents). The resulting number of unique words in our dataset is 63634, the number of POS bigrams is 5864. Since the overall number of unique POS trigrams and fourgrams is prohibitively large, we apply more aggressive term filtering: we consider trigrams that appear in at least 10 documents (44499 trigrams overall) and fourgrams that appear in between 10 and 99 documents (114476 fourgrams overall).

6. Results and discussion

To establish a baseline, we first apply Weka's implementation [32] of the simple k-means algorithm to the BOW and POS bigram representations of the documents. On both representations the algorithm demonstrates poor performance (see Table 1). On POS bigrams it manages to separate the `W_fict_prose` and `W_news_script` genres from the others, putting most of the rest in one large cluster. On the BOW representation it ends up with one cluster of 643 documents (95% of the dataset).

We use Xuerui Wang's implementation [19] of LDA that performs Gibbs sampling with 10000 sampling iterations. For MDC, we use our implementation[2] with 10 random restarts of each optimization iteration.

As we can see from Table 1, MDC achieves more than 50% accuracy with both BOW and POS bigram document representations. Note that a random assignment of documents into clusters would lead to about 5% accuracy on our dataset, so above 50% accuracy is an

2. `http://www.cs.umass.edu/~ronb/mdc.html`

impressive result for a purely unsupervised method on a large, well-balanced dataset. The LDA+BOW system obtains exactly the same result as MDC+BOW in terms of clustering accuracy, and even slightly higher in terms of Jaccard index. However, LDA demonstrates strictly inferior performance (lower than MDC by 6% absolute) on the POS bigram representation.

We also show in Table 1 that the BOW model significantly outperforms the POS model (by more than 4% absolute). This supports our hypothesis that contextual features (such as words) play a more important role for genre classification than stylistic features (as POS ngrams).

To give some insight on the differences in MDC performance on BOW and POS bigrams, we present Table 2 that shows the distribution of documents of each genre over the generated clusters. For each genre we show a list of sizes (in number of documents) of this genre's representation in various clusters. We sort this list by the size of the representation from the largest to the smallest. An asterisk after the number of documents means that this genre is dominant in the corresponding cluster. A heavy tailed distribution (such as the one of `W_non_ac_soc_science`) implies that the genre is spread over many clusters which is clearly a failure. In contrast, a peaked distribution (e.g., of `W_non_ac_tech_engin`) with an asterisk on its largest component means that the genre was successfully identified.

As we can see from the table, MDC performs similarly on BOW and POS bigrams. However, some significant differences can be found. For example, genres `W_biography`, `W_commerce` and `W_institut_doc` are successfully identified by MDC+BOW but not by MDC+POS, while MDC+POS better recognizes `W_newsp_brdsht_nat_social` and `W_pop_lore`. A 3-way MDC with both BOW and POS that would take advantage of the both approaches may have a good chance to show even better results.

Indeed, we obtain a strong result with the 3-way MDC: 58.5% accuracy. The last column of Table 2 presents the analysis of this result by genre. For many genres (such as `W_non_ac_nat_science`) we enlarge their dominant representations. We also manage to identify four of the five genres that were in disagreement between BOW and POS models (as discussed above). However, we no longer recognize `W_ac_polit_law_edu`, which indicates that the results might potentially be improved even more.

One could argue that the direct comparison of results obtained by the BOW and POS bigram models is actually unfair because the number of BOW features is one order greater than the number of POS bigrams, so that the BOW model naturally outperforms the POS bigram model because it just contains more information. However, this argument cannot be empirically proved. We test MDC with POS trigrams and fourgrams, as well as with POS unigrams, and show that while the MDC performance with unigrams is significantly lower than with bigrams, trigrams and fourgrams do not significantly improve the results of bigrams. In Figure 2(a) we can see that when moving from bigrams to trigrams and fourgrams, the graph has a slightly positive slope, however the results become noisier (the standard error becomes higher) which diminishes statistical significance of the improvement. A conclusion that can be made from this experiment is that the Bag-Of-POS-bigrams model appears to be rich enough to capture genres of documents.

A common belief is that stopwords and other high frequency words can be good features for discrimination of documents by genre (see, e.g. [27]). It is interesting to see whether we can support this hypothesis with empirical evidence. To show this, we conduct the

Genre	MDC with POS bigrams	MDC with BOW	LDA with BOW	MDC with BOW and POS bigrams
W_ac_humanities_arts	9* 6* 6 4 2 2 1 1 1	9* 6* 5 5 3 2 1 1	7 6 5 5 4 4 1	9 6* 5 5 4 1 1 1
W_ac_nat_science	23* 4 2 2 1	24* 6 1 1	12* 11* 9	27* 4 1
W_ac_polit_law_edu	14* 8 5 2 1 1 1	20* 5 2 2 1 1 1	19* 7 4 2	17 6 4 2 1 1 1
W_ac_soc_science	11* 9* 6 5 1	12* 10* 7 1 1 1	12* 9* 8* 1 1 1	16* 7 6 3
W_advert	14* 11 3 2 2	18* 3 3 2 2 1 1 1 1	22* 2 2 2 1 1 1 1	23* 2 1 1 1 1 1 1 1
W_biography	15* 8 6 1 1 1	12 7 6 3 2 1 1	16* 6 4 2 2 1 1	16* 6 6* 2 1 1
W_commerce	10* 5 5 4 2 2 1 1 1 1	13 10 6 1 1 1	16 5 4 2 2 1 1 1	9* 9 4 3 3 2 1 1
W_fict_prose	22* 7 3	25* 6 1	30* 2	24* 6 2
W_institut_doc	15* 6 5 5 1	18 6 4 1 1 1 1	17* 7 4 2 2	14 11* 3 1 1 1 1
W_newsp_brdsht_nat_arts	25* 5 1 1	28* 1 1 1 1	30* 2	27* 2 2 1
W_newsp_brdsht_nat_commerce	26* 2 1 1 1 1	32*	28* 2 1 1	31* 1
W_newsp_brdsht_nat_report	32*	32*	30* 2	32*
W_newsp_brdsht_nat_social	9 7 4 4 2 2 1 1 1 1	11* 6 4 3 2 2 1 1 1 1	10 7 6 3 2 1 1 1 1	14 6 3 2 2 2 1 1 1
W_news_script	32*	32*	31* 1	32*
W_non_ac_humanities_arts	11* 8 3 2 2 2 1 1 1 1	9* 6 5 3 3 2 2 2	10* 7 5 3 2 2 2 1	14* 5 3 3 2 1 1 1 1 1
W_non_ac_nat_science	14* 5* 3 2 2 2 1 1 1 1	18* 11 2 1	11* 9 7 2 2 1	29* 1 1 1
W_non_ac_polit_law_edu	11* 4 4 3 3 2 2 1 1 1	11 10* 5 3 2 1	10* 10* 3 3 2 2 1 1	10* 6 5 5 2 2 1 1
W_non_ac_soc_science	5 5 4 3 3 2 2 2 2 1 1 1 1	7 5 4 4 3 2 2 2 2 1	7 6 5 5 3 2 1 1 1 1	5 5 4 3 3 3 2 2 2 1 1 1
W_non_ac_tech_engin	32*	32*	32*	32*
W_pop_lore	11 6 6 5 4	10* 9* 4 4 2 2 1	12 8 6 3 2 1	16* 8 3 2 2 1
W_religion	11* 5 4 4 2 2 1 1 1 1	18* 6 2 1 1 1 1 1 1	20* 6* 2 1 1 1 1	18* 6* 3 1 1 1 1 1

Table 2: Performance of various methods by genre. For each genre we show a list of sizes (in number of documents) of this genre's representation in various clusters. We sort this list by the size of the representation from the largest to the smallest. An asterisk after the number of documents means that this genre is dominant in the corresponding cluster.

following experiment. We put various thresholds on the low frequency words in the BOW representation of the documents. We consider four such thresholds: our initial setup, when we filter out words that appear in less than 3 documents, as well as three new ones: 10, 20 and 50 documents. Note that the new thresholds and especially the most restrictive one (50) leave us with highly frequent words only: since our dataset consists of 672 documents, filtering out words that appear in less than 50 documents causes removal of over 93% of unique words from the dataset. We run MDC on the four representations. Figure 2(b) shows results of this experiment. We can see that although the graph has a negative slope, the decrease in the results is insignificant. With 7% of words from the original dataset the MDC system obtains only 2.5% lower accuracy than with 38% of words (where the rest appear in only one or two documents and can be removed with high confidence). This result confirms that high frequency words are important for genre classification.

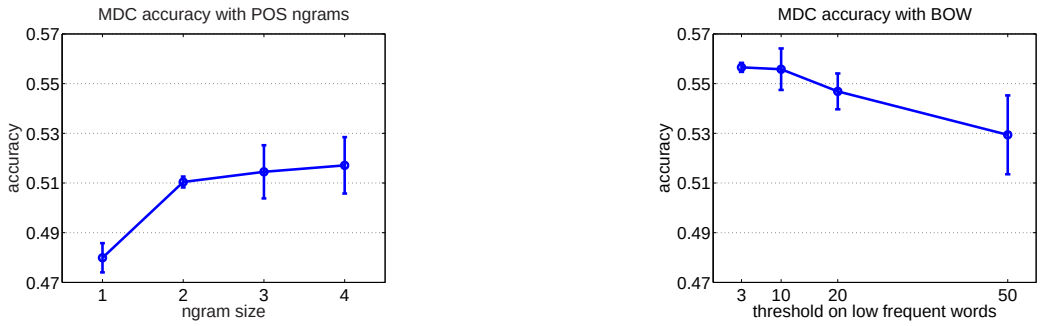

Figure 2: MDC accuracy as a function of: (a) the size of POS ngram (i.e. 1-grams, 2-grams, 3-grams and 4-grams); (b) threshold on a low frequency words – a point i on the X axis means that in this experiment words that appear in less than i documents are excluded from the consideration.

7. Conclusion and future work

The goal of this paper is to establish a framework for unsupervised non-topical classification of text and to illustrate an application of this framework to a specific task of clustering by genre. We extensively study all aspects of this task, including feature selection / construction possibilities, application of various clustering algorithms and exploitation of the multi-modal nature of the data. We infer that (a) contextual features (such as words) are more appropriate for the problem of clustering by genre than are stylistic features (such as Part-Of-Speech ngrams); (b) the Multi-way Distributional Clustering method [4] is a good choice for genre clustering, not only because it obtains decent empirical results but also because it allows the use of rich multidimensional document representations without manually crafting a compelling set of features.

In the future, we will explore other types of document representations and learning algorithms for the problem of genre clustering. One of the natural choice would be to represent documents as Bag-Of-Word-Ngrams, in analogy to the Bag-Of-POS-Ngrams discussed in this paper. However, our previous experience indicates that word ngrams rarely help to improve text classification results (see, e.g., [3]). We are also planning to apply the proposed framework to other non-topical clustering problems, such as clustering Blog postings by the author's mood. Recently, Mishne [21] proposed a method for (supervised) text classification by mood and released a large dataset with mood labels given by the Blog authors themselves. We intend to approach the problem of clustering by mood and to test it on this data. Our pilot research shows that the Bag-Of-Words document representation cannot be used for this task because the vocabulary of Blog postings almost never correlates with the author's current mood. Some non-contextual features should be used, such as punctation, emoticons (smilies) etc. We have obtained promising preliminary results with MDC on the Blog data.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by the Ministry of Education, Culture, Sports, Science and Technology of Japan under grant number KAKENHI-17680011.

References

[1] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3), Aug. 2003.

[2] L. D. Baker and A. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st International ACM SIGIR Conference*, pages 96–103, 1998.

[3] R. Bekkerman and J. Allan. Using Bigrams in Text Categorization. CIIR Technical Report, University of Massachusetts at Amherst, 2004.

[4] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 41–48, 2005.

[5] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. On feature distributional clustering for text categorization. In *Proceedings of the 24th Intl. ACM SIGIR Conference*, pages 146–153, 2001.

[6] D. Biber. *Variation Across Speech and Writing.* Cambridge University Press, 1988.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pages 993-1022, 2003.

[8] L. Burnard. User Reference Guide for the British National Corpus. Technical report, Oxford University Computing Services, 2000.

[9] A. Finn, N. Kushmerick, and B. Smyth. Genre classification and domain transfer for information filtering. In *Advances in Information Retrieval*, LNCS 2291. Springer-Verlag, 2002.

[10] P. Jaccard. The distribution of flora in the alpine zone. *New Phytologist.* 11, pages 37–50, 1912.

[11] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, 1988.

[12] J. Karlgren and D. R. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 1071–1075, 1994.

[13] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, 1997.

[14] M. Koppel and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[15] G. Kumaran, R. Jones, and O. Madani. Biasing web search results for topic familiarity. In *Proceedings of the ACM 14th Conference on Information and Knowledge Management*, pages 271–272, 2005.

[16] D. Y.-W. Lee. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), pages 37–72, 2001.

[17] Y.-B. Lee and S. H. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th International ACM SIGIR Conference*, pages 145–150, 2002.

[18] R. A. J. Matthews and T. V. N. Merriam. Neural computation in stylometry i: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8(4):203–209, 1993.

[19] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and Role Discovery in Social Networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 786–791, 2005.

[20] M. Meila. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd International Conference on Machine learning*, pages 577–584, 2005.

[21] G. Mishne. Experiments with mood classification in Blog posts. In *Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access*, 2005.

[22] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

[23] A. Rauber and A. Müller-Kögler. Integrating automatic genre analysis into digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10, 2001.

[24] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, 2004.

[25] M. Santini. A Shallow Approach to Syntactic Feature Extraction for Genre Classification. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, 2004.

[26] Y. Seki, K. Eguchi, and N. Kando. Analysis of multi-document viewpoint summarization using multi-dimensional genres. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 150–153, 2004.

[27] E. Stamatatos, N. Fakotakis, and G. K. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 808–814, 2000.

[28] J. M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge University Press, 1990.

[29] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[30] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.

[31] P. Willet. Recent trends in hierarchical document clustering: a critical review. *Information Processing and Management*, 24(5), pages 577–597, 1988.

[32] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* 2nd edition, Morgan Kaufmann, 2005.